Universidad del Valle de Guatemala
Data Science - CC3066 - Sección 20
Catedrático: LUIS ROBERTO FURLAN COLLVER

Integrantes:
JULIO ROBERTO HERRERA SABAN
GUIDO SEBASTIAN PADILLA ALDANA
OSCAR ANDRE PAREDEZ URIZAR
DIEGO DE JESUS ARREDONDO TURCIOS

Laboratorio 2 - Series de tiempo

Link del Repositorio: https://github.com/jurhs2000/data-science-lab2

Análisis exploratorio

Para el análisis exploratorio empezamos con un vistazo rápido a las primeras 5 filas del *dataset*, con esto vemos las columnas existentes y el tipo de dato que contienen.

	dt	AverageTemperature	AverageTemperatureUncertainty	Country
0	1743-11-01	4.384	2.294	Åland
1	1743-12-01	NaN	NaN	Åland
2	1744-01-01	NaN	NaN	Åland
3	1744-02-01	NaN	NaN	Åland
4	1744-03-01	NaN	NaN	Åland

Figura 1: Primeras 5 filas del *dataset*

Este vistazo rápido nos da indicios de que pueden existir muchos valores *NaN* así que nos apoyamos de *Profile Report* para realizar un reporte más completo de las 4 variables, que representan 577462 de observaciones.

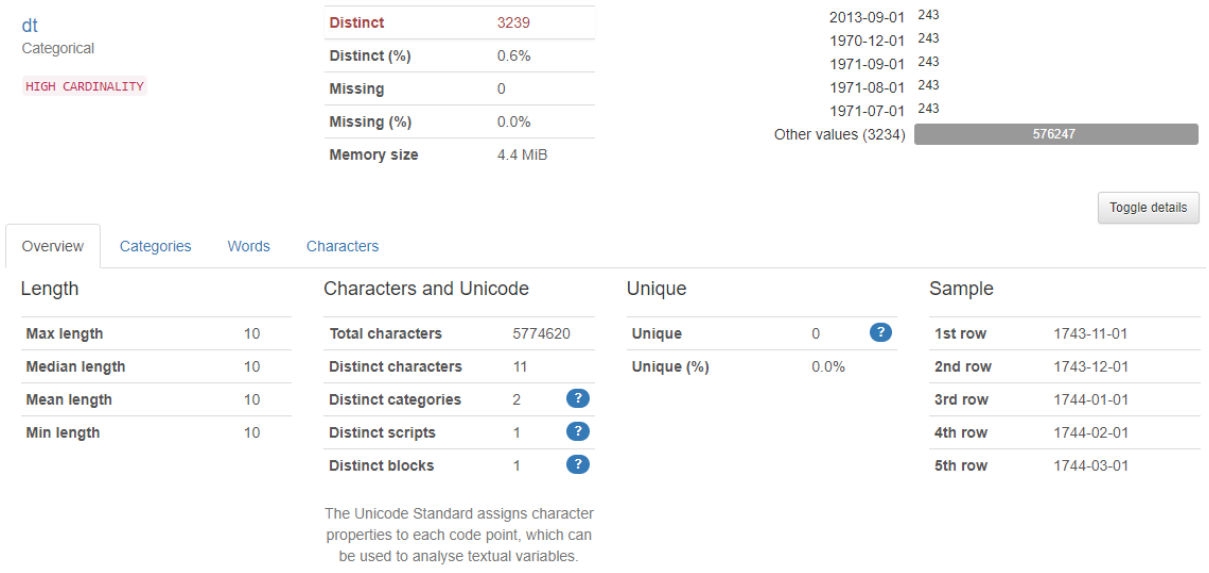


Figura 2: Reporte de la variable dt (date).

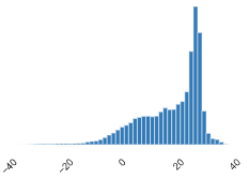
AverageTemperature

Real number (ℝ)

MISSING

Distinct	71085
Distinct (%)	13.0%
Missing	32651
Missing (%)	5.7%
Infinite	0
Infinite (%)	0.0%
Mean	17.19335423

Minimum	-37.658
Maximum	38.842
Zeros	7
Zeros (%)	< 0.1%
Negative	44864
Negative (%)	7.8%
Memory size	4.4 MiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	-37.658
5-th percentile	-3.267
Q1	10.025
median	20.901
Q3	25.814
95-th percentile	28.423
Maximum	38.842
Range	76.5
Interquartile range (IQR)	15.789

Descriptive statistics

Standard deviation	10.95396645
Coefficient of variation (CV)	0.6371046799
Kurtosis	1.068769602
Mean	17.19335423
Median Absolute Deviation (MAD)	5.982
Skewness	-1.114296766
Sum	9367128.513
Variance	119.9893809
Monotonicity	Not monotonic

Figura 2: Reporte de la variable AverageTemperature

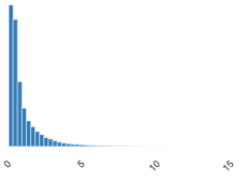
AverageTemperatureUnce...

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct	8979
Distinct (%)	1.6%
Missing	31912
Missing (%)	5.5%
Infinite	0
Infinite (%)	0.0%
Mean	1.0190569

Minimum	0.052
Maximum	15.003
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	4.4 MiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0.052
5-th percentile	0.187
Q1	0.323
median	0.571
Q3	1.206
95-th percentile	3.337
Maximum	15.003
Range	14.951
Interquartile range (IQR)	0.883

Descriptive statistics

Standard deviation	1.201930387
Coefficient of variation (CV)	1.179453656
Kurtosis	13.86904492
Mean	1.0190569
Median Absolute Deviation (MAD)	0.309
Skewness	3.142598416
Sum	555946.492
Variance	1.444636654
Monotonicity	Not monotonic

Figura 3: Reporte de la variable AverageTemperatureUncertainty

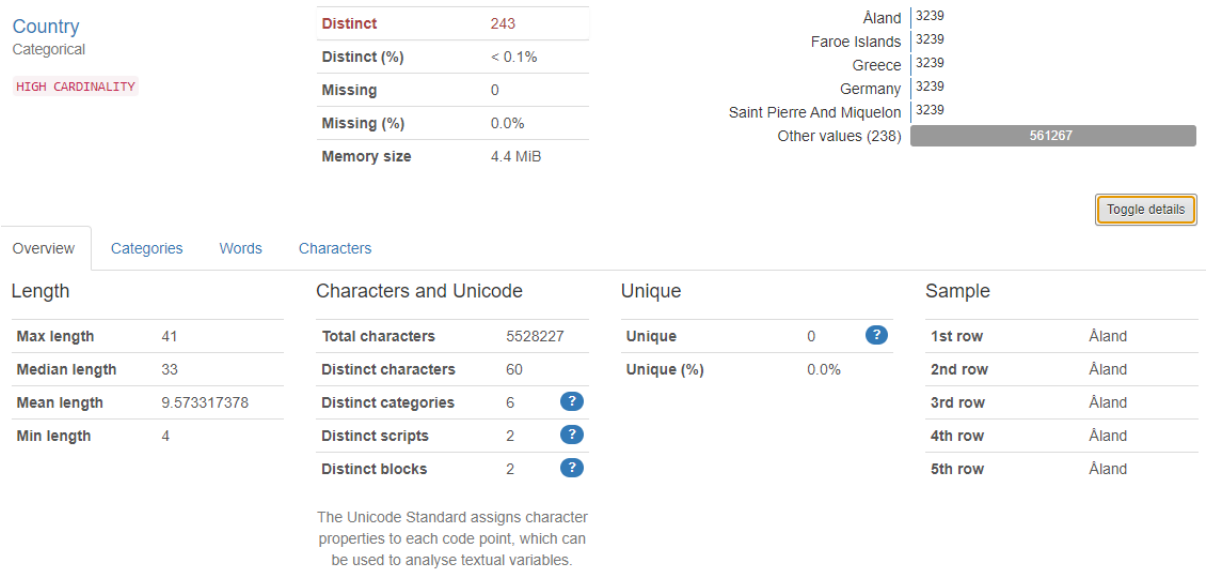


Figura 4: Reporte de la variable Country

A partir de las estadísticas y gráficas dadas por el reporte vemos la alta cardinalidad de las variables dt y country, así como que las variables AverageTemperature y AverageTemperatureUncertainty no tienen correlación y solo llegan a tener hasta un 5.7% de valores faltantes por lo que procedemos a limpiar el *dataset* quitando los valores faltantes. Con esto vemos que nos quedan los 577462 valores y que los valores de *dt* van de la fecha 1743-11-01 a 2013-09-01.

	dtypes	count	null_sum	null_pct	nunique	min	25%	50%	75%	max	mean	median	std	skew
AverageTemperature	float64	577462	0	0.0	71085	-37.658	10.354	21.271	25.777	38.842	17.399923	21.271	10.83239	-1.146609
AverageTemperatureUncertainty	float64	577462	0	0.0	8979	0.052	0.333	0.617	1.34	15.003	1.077051	0.617	1.218259	2.886287
Country	object	577462	0	0.0	243	Afghanistan	-	-	-	Åland	-	-	-	-
dt	object	577462	0	0.0	3239	1743-11-01	-	-	-	2013-09-01	-	-	-	-

Figura 5: Resumen exploratorio de las estadísticas con el *dataset* limpio.

Análisis de la serie de tiempo

Habiendo limpiado los datos se analiza la serie de tiempo a través de varias gráficas que permiten observar el comportamiento de la temperatura promedio. El análisis incluye la observación de los meses y la temperatura promedio en esos meses durante toda la serie de tiempo; también un acercamiento al comportamiento de la temperatura promedio durante los últimos 20 años; un análisis simple a los valores de temperatura promedio y la descomposición de la serie en componentes.

Es notorio que la serie presenta estacionalidad y podemos visualizarlo debido a sus patrón de comportamiento el cual tiende a subir y bajar de una forma peculiar y notoria, siendo presente de forma mensual lo cual implica que efectivamente tiende a variar según el mes, siendo acorde a las estaciones del planeta donde las temperaturas tienden a subir y bajar dependiendo de lo mismo. En cuanto a la tendencia podemos denotar que en los últimos se hace presente de forma significativa y este tiene una tendencia creciente a lo largo del tiempo, significando un aumento en la temperatura en general, a grandes rasgos un aumento en la temperatura por cada año que transcurre.

Análisis de temperatura promedio por mes

Esta gráfica muestra la temperatura promedio en cada mes, tomando en cuenta los meses de toda la serie de tiempo. Esto nos muestra qué meses tienen una mayor y menor temperatura promedio y es coherente que los primeros y los últimos meses del año compartan temperatura promedio. Podemos observar que los primeros y últimos meses tienen menor temperatura promedio (12 ºC) mientras que los meses de mayo a septiembre tienen las mayores temperaturas, siendo julio el mes con mayor temperatura promedio.

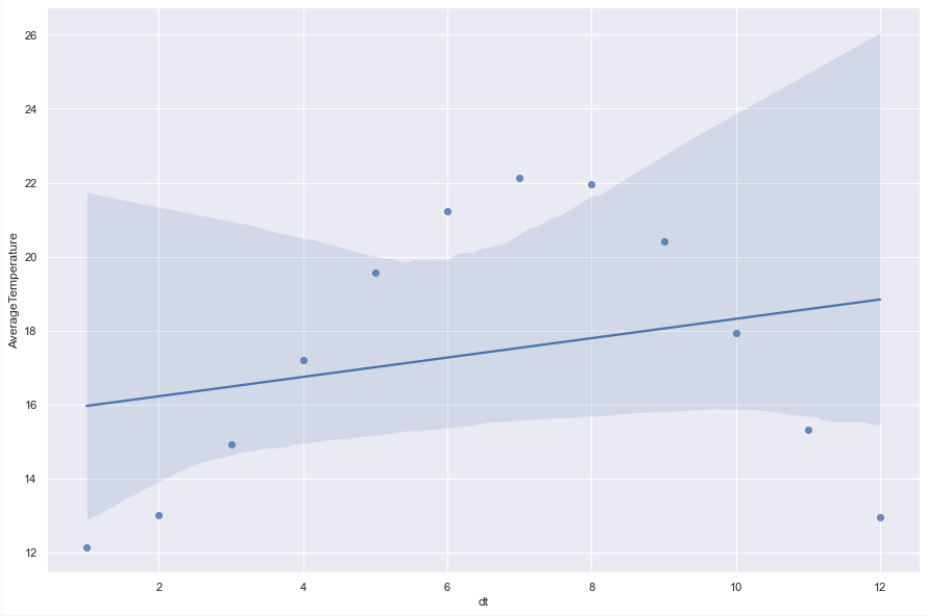


Figura 6: Meses vs. Temperatura promedio

Análisis de temperatura promedio en los últimos años

La serie de tiempo contiene datos históricos desde 1743 por lo que es de ayuda ver el comportamiento durante un periodo de tiempo más corto, precisamente de los últimos 18 años (de la serie de tiempo) ya que si no se aprecia un rango muy alto de valores en la gráfica. Para ello se selecciona el periodo de tiempo y se realiza la gráfica de dispersión con regresión.

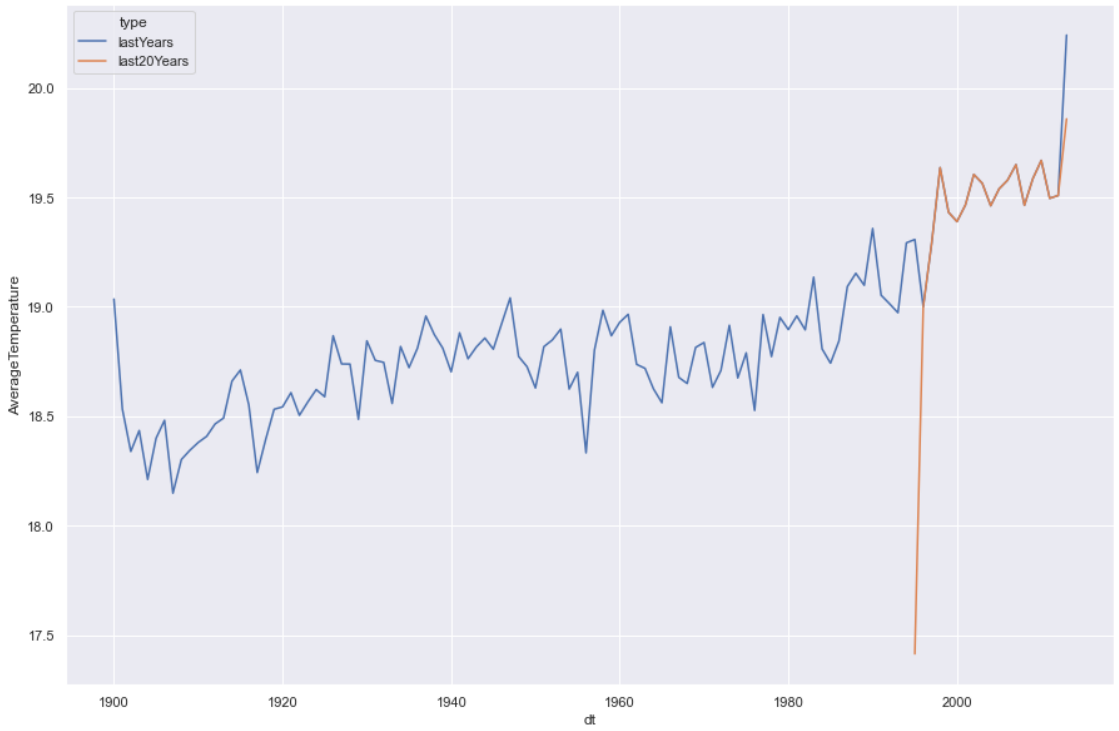


Figura 7: Períodos de tiempo, últimos años desde 1900 y últimos 18 años.

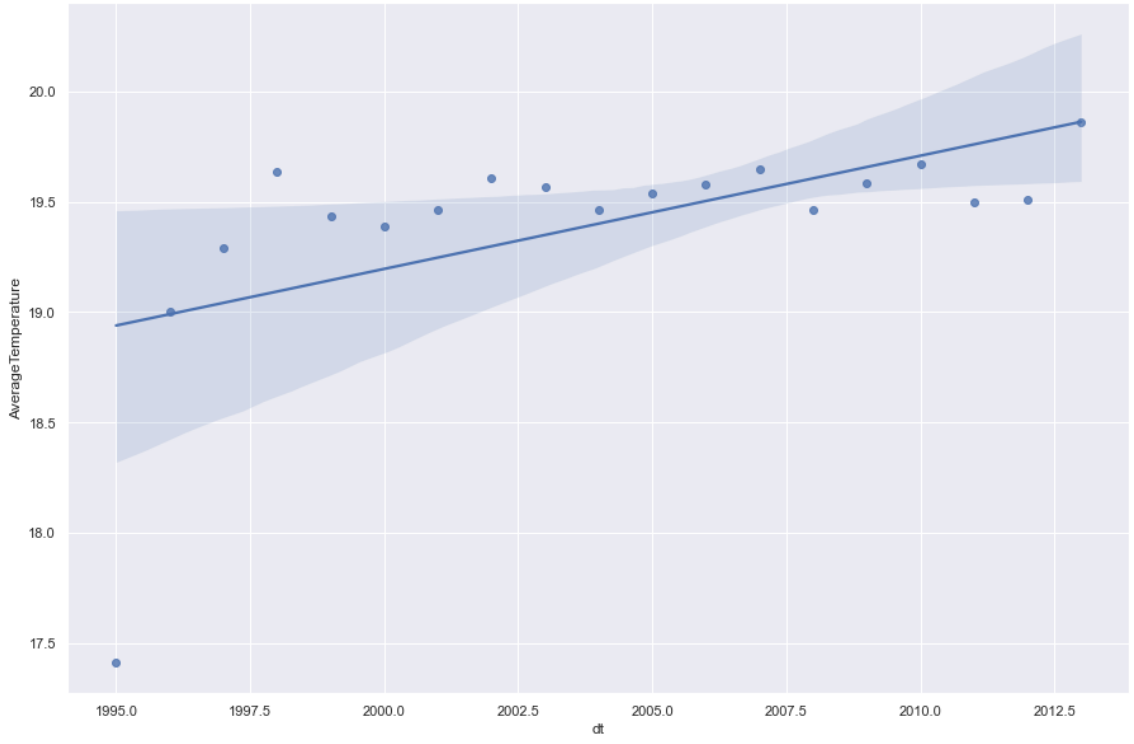


Figura 8: Temperatura promedio en los últimos 20 años (1994 - 2014).

Análisis de valores de temperatura promedio

Se elaboró un histograma de los valores de la temperatura en el *dataset*, lo que nos permite ver que existe una tendencia a las temperaturas entre 20 ºC y 30 ºC. Y que existen datos de temperaturas que llegan hasta un mínimo de -37 ºC y a un máximo de 38 ºC.

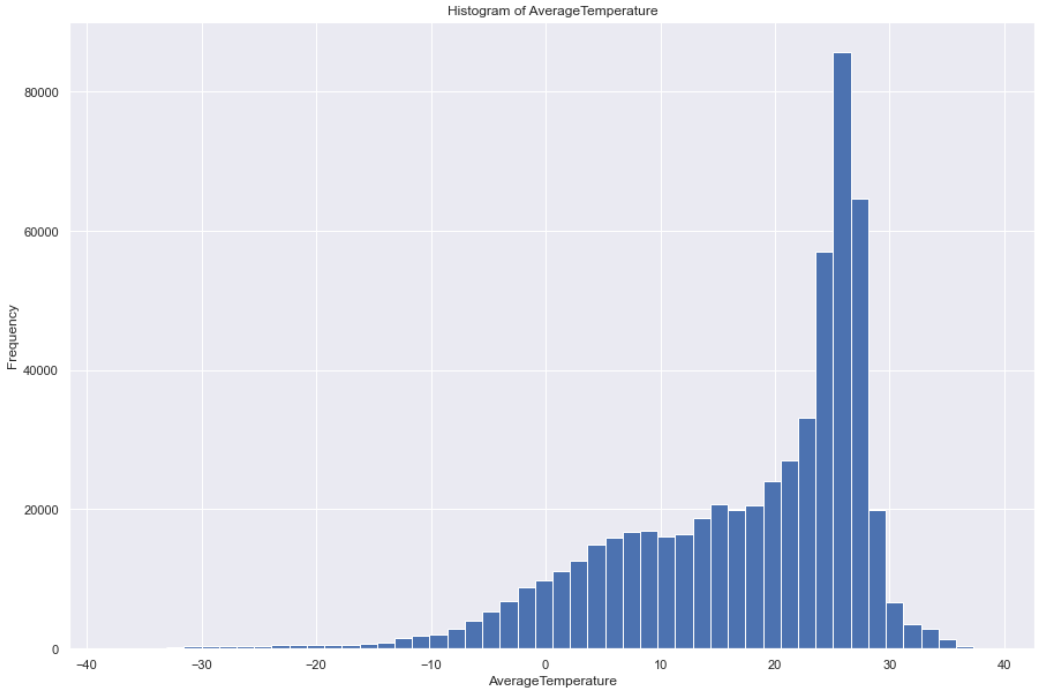


Figura 9: Histograma de la temperatura promedio

Descomposición de la serie de tiempo en componentes

La serie de tiempo contiene 3 componentes: Tendencia, estacional y aleatorio. Estos se pueden obtener usando la librería *statsmodel*, indicando la serie de tiempo a utilizar y el periodo.

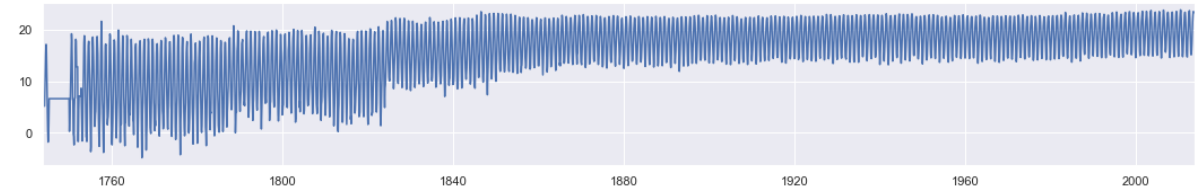


Figura 10: Promedio de Temperaturas promedio en el *dataset*.

Componente Tendencia

El componente tendencia indica un cambio a largo plazo a partir del nivel medio. Esta se identifica con movimientos leves a largo plazo, por lo que se aprecia que este componente en la serie de tiempo muestra una tendencia de aumento.

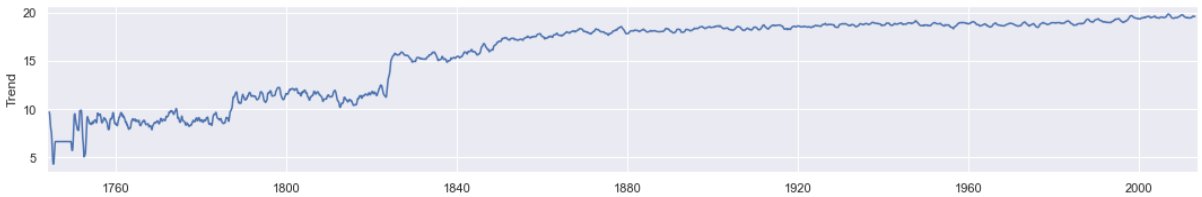


Figura 11: Componente Tendencia de la serie de tiempo.

Componente Estacional

El componente estacional indica la variación periódica en el *dataset* ya sin el componente tendencia, es por eso que vemos que este componente oscila siempre ya que es coherente con el ciclo de temperaturas en cada año.

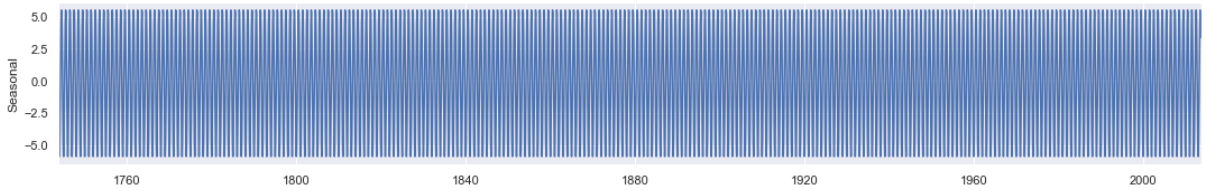


Figura 12: Componente Estacional de la serie de tiempo.

Componente Aleatorio

El componente aleatorio no responde a ningún patrón y nos indica que la serie de tiempo contiene influencias aleatorias, en este caso el modelo aditivo usado es conveniente ya que se obtuvo un componente aleatorio (ya que si se observa un patrón en este componente significa que no es correcto el modelo). Suponiendo sobre el *dataset*, se puede decir que los primeros años muestra mayor aleatoriedad ya que los valores al ser más antiguos pudieron haber sido recogidos de distintas fuentes.

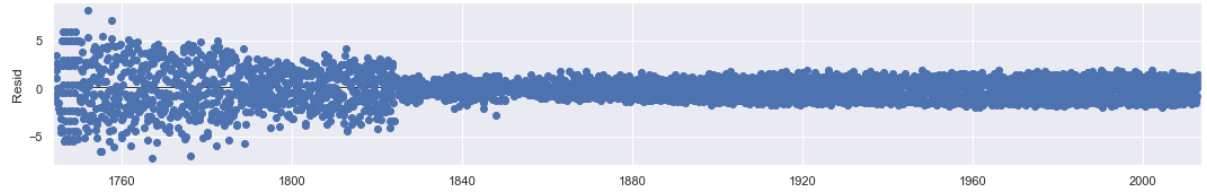


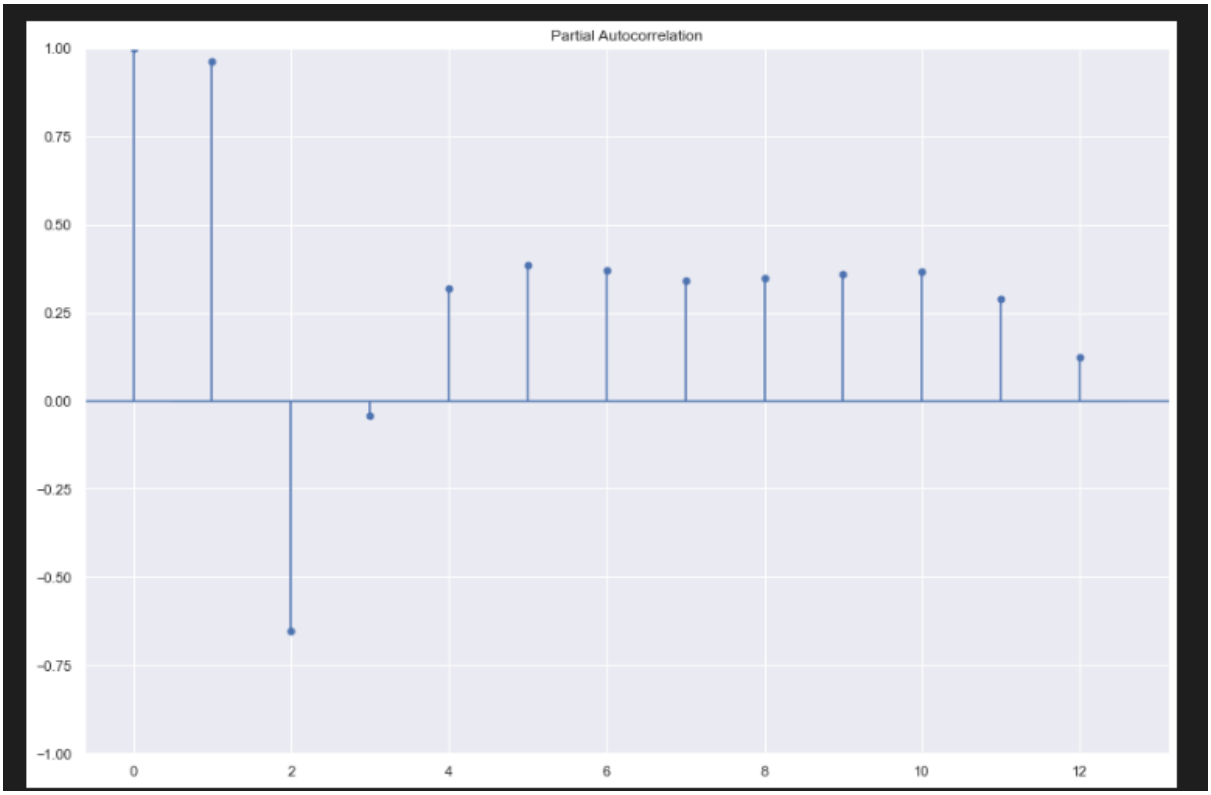
Figura 13: Componente Aleatorio de la serie de tiempo

Generación de modelos

Modelo SARIMA

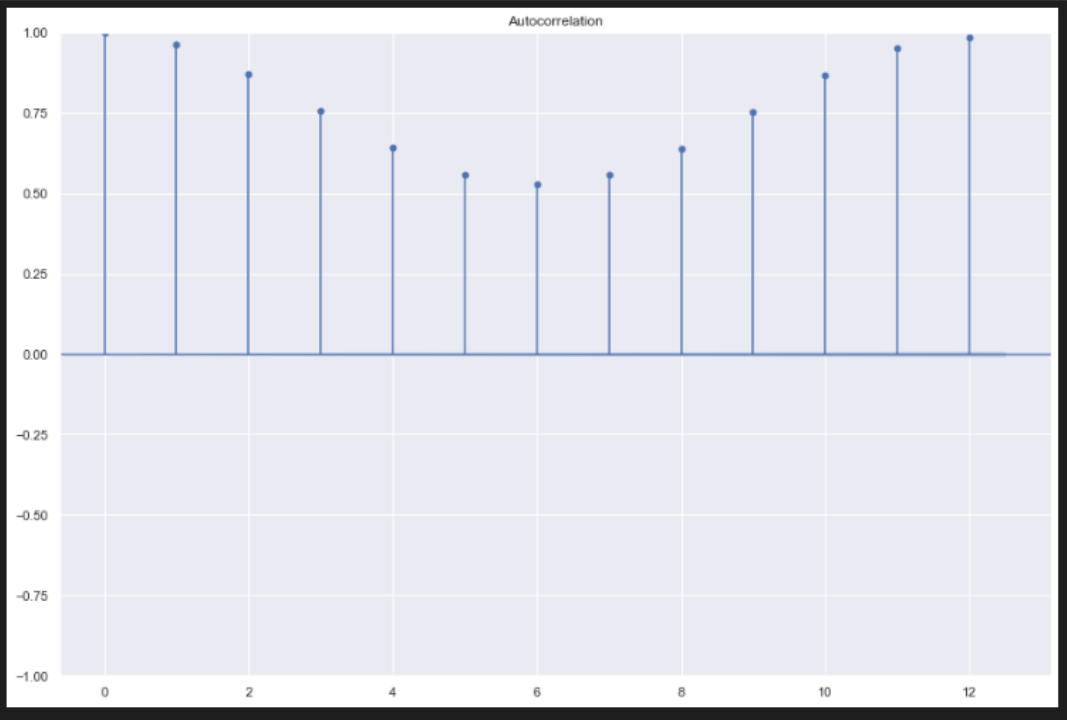
```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
plot_pacf(seasonal, lags =12)
plt.show()
```

✓ 0.4s




```
plot_acf(seasonal, lags =12)
plt.show()
```

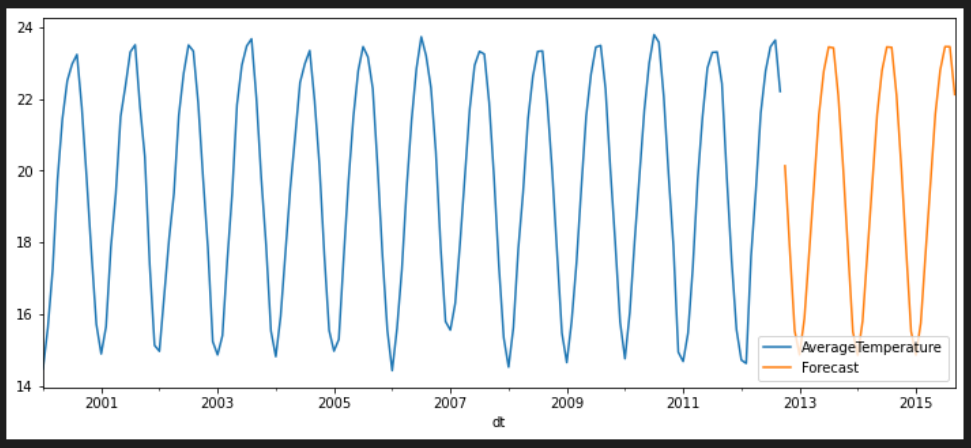
[16] ✓ 53.5s



```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                 AverageTemperature   No. Observations:                  153
Model:            SARIMAX(1, 1, 1)x(2, 1, 1, 12)   Log Likelihood                -23.023
Date:                           Mon, 08 Aug 2022   AIC                             58.045
Time:                                   13:27:43   BIC                             75.695
Sample:                             01-01-2000   HQIC                            65.217
                                  - 09-01-2012
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2896      0.119      2.441      0.015       0.057       0.522
ma.L1         -0.9946      0.338     -2.946      0.003      -1.656      -0.333
ar.S.L12      -0.1053      0.177     -0.594      0.553      -0.453       0.242
ar.S.L24      -0.0534      0.130     -0.411      0.681      -0.309       0.202
ma.S.L12      -0.9981     17.138     -0.058      0.954     -34.589      32.593
sigma2         0.0611      1.033      0.059      0.953      -1.965       2.087
===================================================================================
Ljung-Box (L1) (Q):                   0.57   Jarque-Bera (JB):                40.29
Prob(Q):                              0.45   Prob(JB):                         0.00
Heteroskedasticity (H):               1.12   Skew:                            -0.80
Prob(H) (two-sided):                  0.70   Kurtosis:                         5.08
===================================================================================
...
```

```
<AxesSubplot:xlabel='dt'>
```

</>



```
No frequency information was provided, so inferred frequency MS will be used.
```

SARIMAX Results

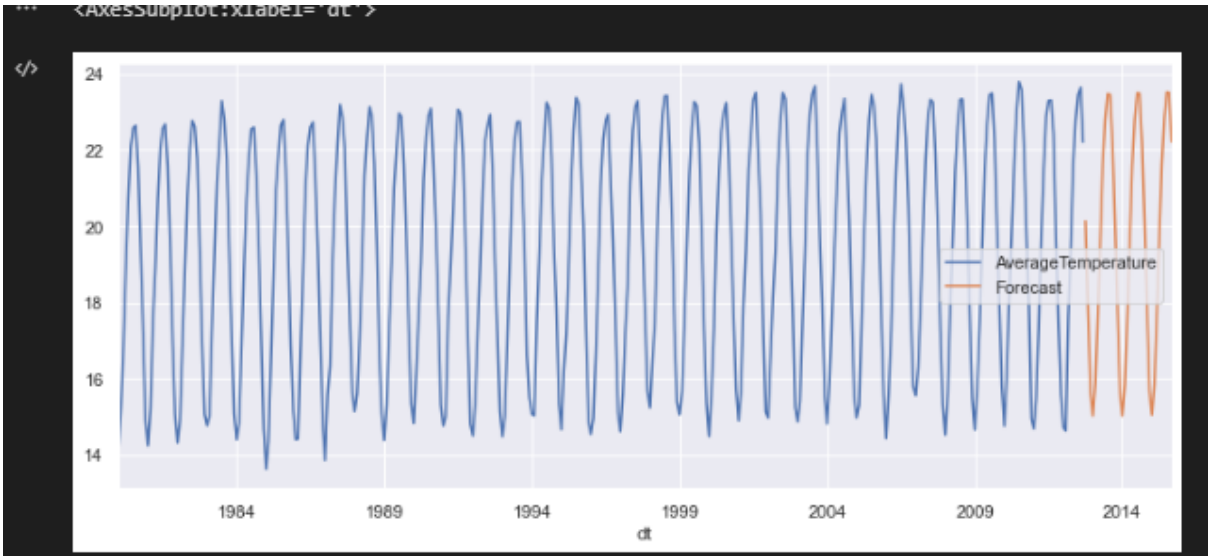
Dep. Variable:	AverageTemperature	No. Observations:	393
Model:	SARIMAX(2, 1, 1)x(2, 1, 1, 12)	Log Likelihood	-43.700
Date:	Mon, 08 Aug 2022	AIC	101.400
Time:	15:10:49	BIC	128.981
Sample:	01-01-1980	HQIC	112.345
	- 09-01-2012		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2912	0.049	5.893	0.000	0.194	0.388
ar.L2	0.1672	0.052	3.205	0.001	0.065	0.269
ma.L1	-0.9711	0.024	-40.587	0.000	-1.018	-0.924
ar.S.L12	-0.0500	0.059	-0.849	0.396	-0.165	0.065
ar.S.L24	-0.0027	0.058	-0.047	0.963	-0.116	0.111
ma.S.L12	-0.9896	0.196	-5.038	0.000	-1.375	-0.605
sigma2	0.0657	0.012	5.341	0.000	0.042	0.090

Ljung-Box (L1) (Q):	0.24	Jarque-Bera (JB):	37.69
Prob(Q):	0.62	Prob(JB):	0.00
Heteroskedasticity (H):	0.85	Skew:	-0.00
Prob(H) (two-sided):	0.37	Kurtosis:	4.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```
<AxesSubplot:xlabel='dt'>
```



El modelo SARIMA es un modelo paramétrico que trata de obtener la representación de la serie en términos de la interrelación temporal de sus elementos. El test de diferenciación

una vez por eso se tomó parámetro 1, en la autocorrelación parcial luego de 2 la desigualdad se detiene, y p es 1 debido a que nunca se detiene en 0 en la autocorrelación. La predicción hecha por sarima con los datos fue la correcta debido a que posee una correlación correcta a los datos reales, sin embargo, en los datos reales la tendencia era creciente lo cual no fue predicha por el modelo de sarima.

Modelo Prophet

Prophet se utiliza como un procedimiento de previsión de datos para series de tiempo, que se basa en un modelo aditivo en el que las tendencias no lineales se ajustan a la estacionalidad anual. semanal y diaria (incluyendo etapa de vacaciones).

Modelo Prophet, últimos 3 años. Predicción y comparación

Para este modelo, se quitaron los últimos 3 años de la serie de tiempo y se ajustó el modelo con esta *data*, utilizando un 95% de intervalo de incertidumbre. Se realizaron las predicciones y se compararon con los valores reales del *dataset* de esos últimos 3 años.

Como se puede ver a continuación, el modelo de Prophet nos indica la variabilidad "constante" de temperatura a lo largo de los años. Es decir, se puede visualizar como la temperatura es baja al inicio del año, luego comienza a subir, y más o menos a medio año empieza a bajar a donde inició al inicio del año. Se puede observar que "yhat" (datos resultantes de la predicción) predice bastante bien este comportamiento, ya que sigue el patrón de la "y_true" (datos reales del *dataset*).

Como indicador de desempeño se utilizó el MAPE (Mean Absolute Percentage Error) o Error Absoluto Medio Porcentual el cual nos dice el tamaño del error absoluto en términos porcentuales, este se obtuvo de la librería *sklearn.metrics* y se obtuvo un valor de 0.0224 es decir un 2.24% de error, el cuál es bastante aceptable.

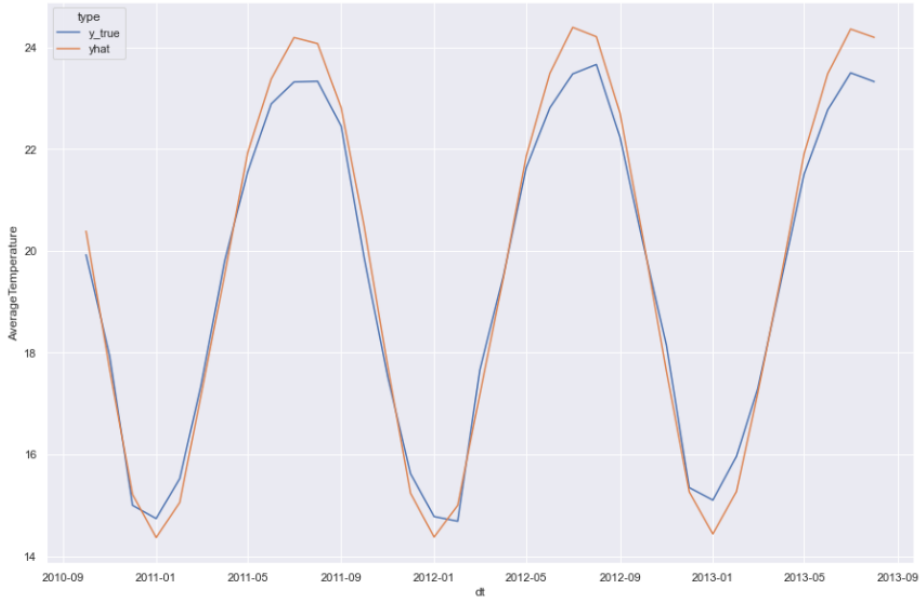


Figura: Comparación de los valores predichos y reales de los últimos 3 años.

Asimismo, se puede observar a continuación que la predicción para la temperatura utilizando y_hat lower y upper siguen bastante bien a los valores originales.

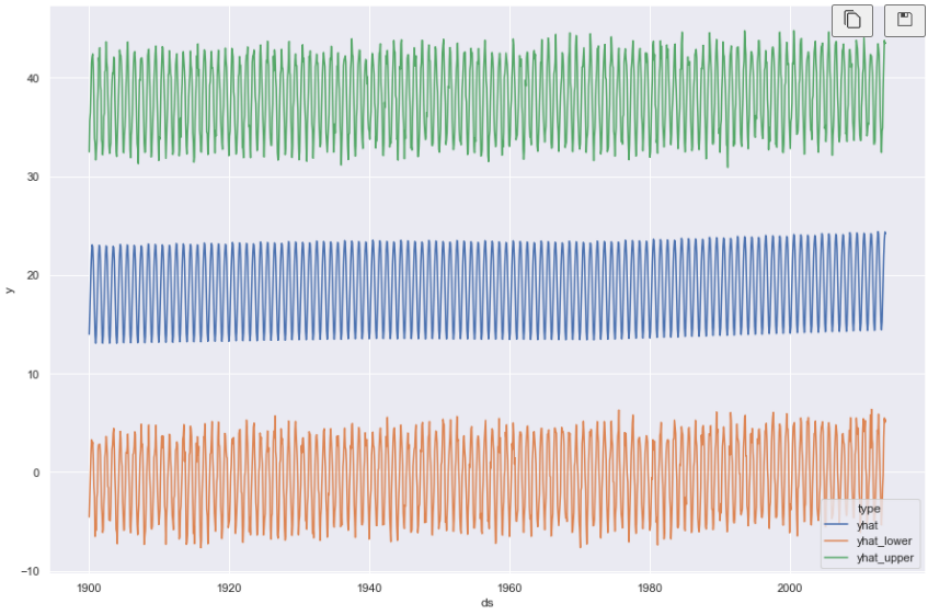


Figura: *yhat_lower*, *yhat* y *yhat_upper* de los años desde 1900.

Finalmente, la última gráfica que nos provee Prophet demuestra el incremento de la temperatura con el pasar de los años, a largo plazo. Se puede ver que ha habido un incremento de la temperatura, que a pesar de que dicho incremento es poco, está presente. Esta gráfica al incluir todos los datos desde el año 1753 es poco visible.

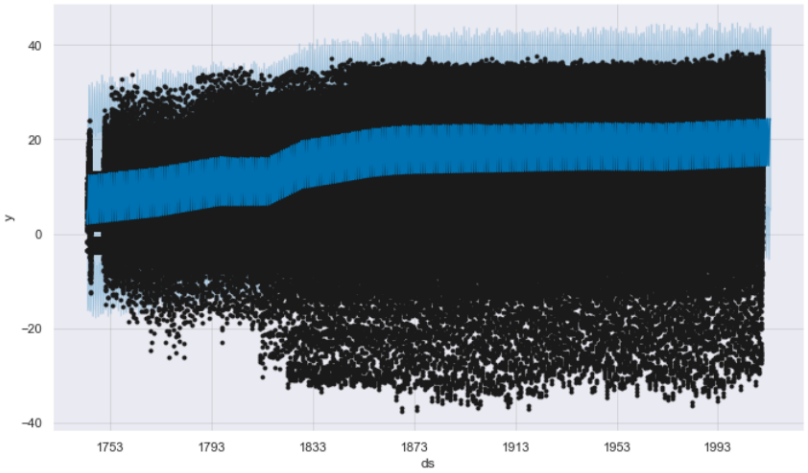
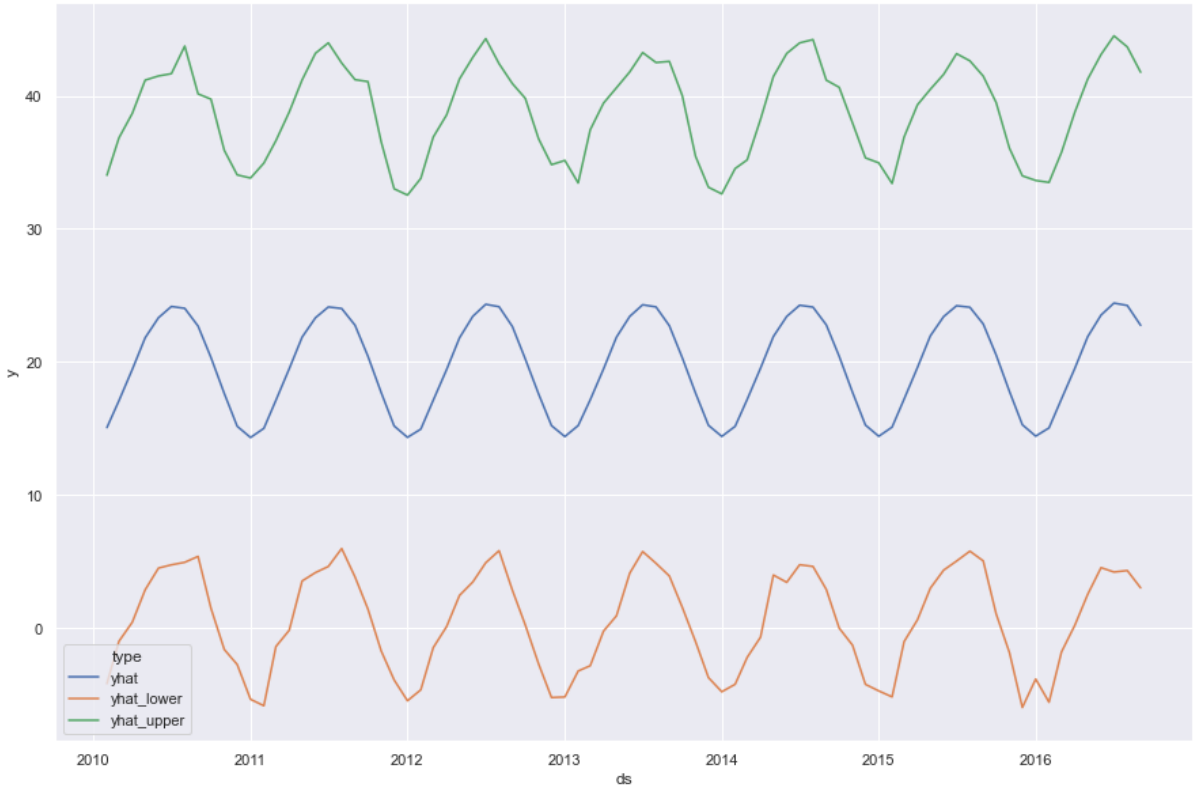


Figura: Gráfica con predicción de todos los años dada por Prophet.

Modelo Prophet 3 años a Futuro

Para el modelo de Prophet de 3 años a futuro, se ajustó el modelo ahora utilizando todo el *dataset*, con un 95% de intervalo de incertidumbre. También se observan que sus valores para yhat, yhat_lower y yhat_upper, siguen bastante acordes a los datos reales, lo cual indica que las predicciones están bastante cerca de las reales, considerando que en el año 2014, 2015 y 2016, la temperatura aumentó realmente en unos 0.69ºC, 1.33ºC y 0.94ºC respectivamente según el National Centers for Environmental Information.


40
30
20
10
0
y
type
yhat
yhat_lower
yhat_upper
2010
2011
2012
2013
2014
2015
2016
ds

Predicción para Guatemala

Análisis de temperatura promedio por mes

Podemos observar en la Figura #8, que los meses en los cuales Guatemala representa una mayor temperatura, de un promedio de 25°, es entre los meses 4 y 6 (abril a junio) y que en el último y primer mes del año, es donde las temperaturas bajan hasta un promedio de 21°.

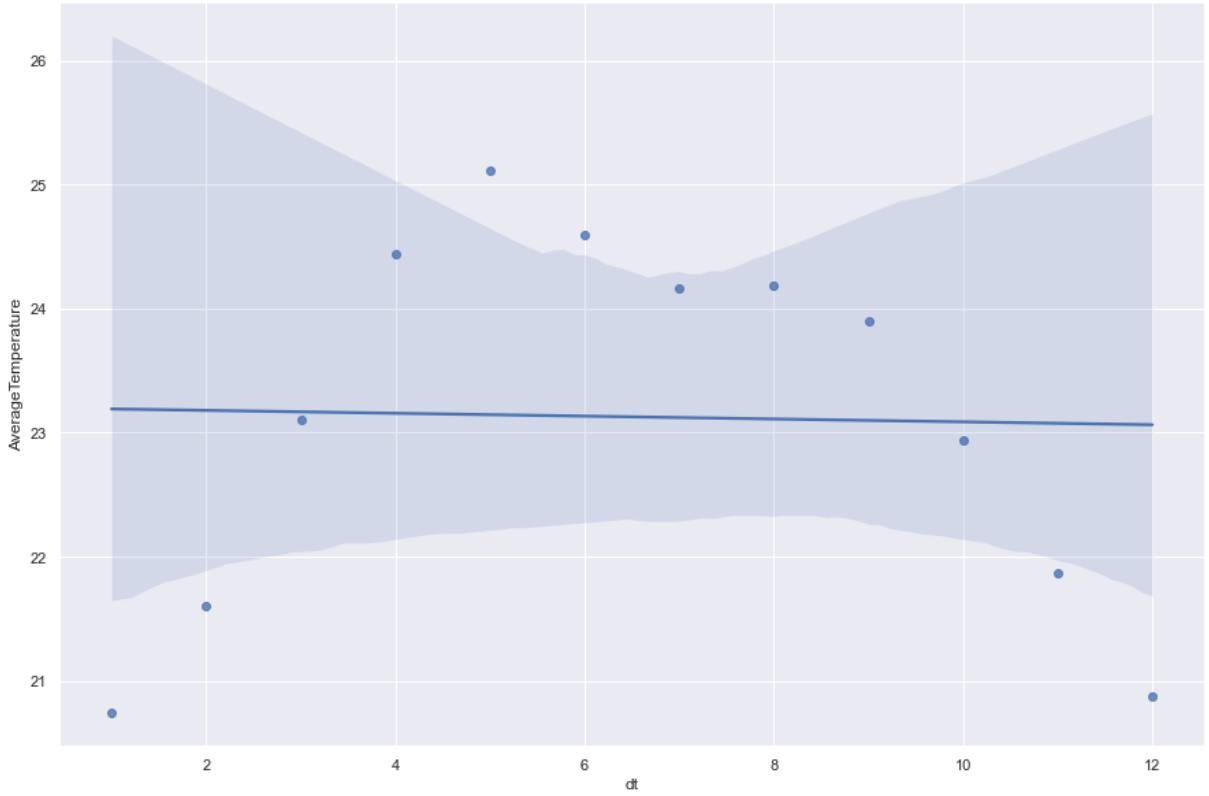


Figura 8: Meses vs. Temperatura promedio en Guatemala

Análisis de temperatura promedio en los últimos años

Tomando los datos históricos desde 1743 nos da una visión general histórica de la temperatura, pero para ver si el calentamiento es un problema, es de ayuda ver el comportamiento durante un periodo de tiempo más corto, precisamente de los últimos 18 años (de la serie de tiempo) ya que si no se aprecia un rango muy alto de valores en la gráfica. Para ello se selecciona el periodo de tiempo y se realiza la gráfica de dispersión con regresión, la cual presenta una recta con tendencia hacia arriba y también con puntos atípicos.

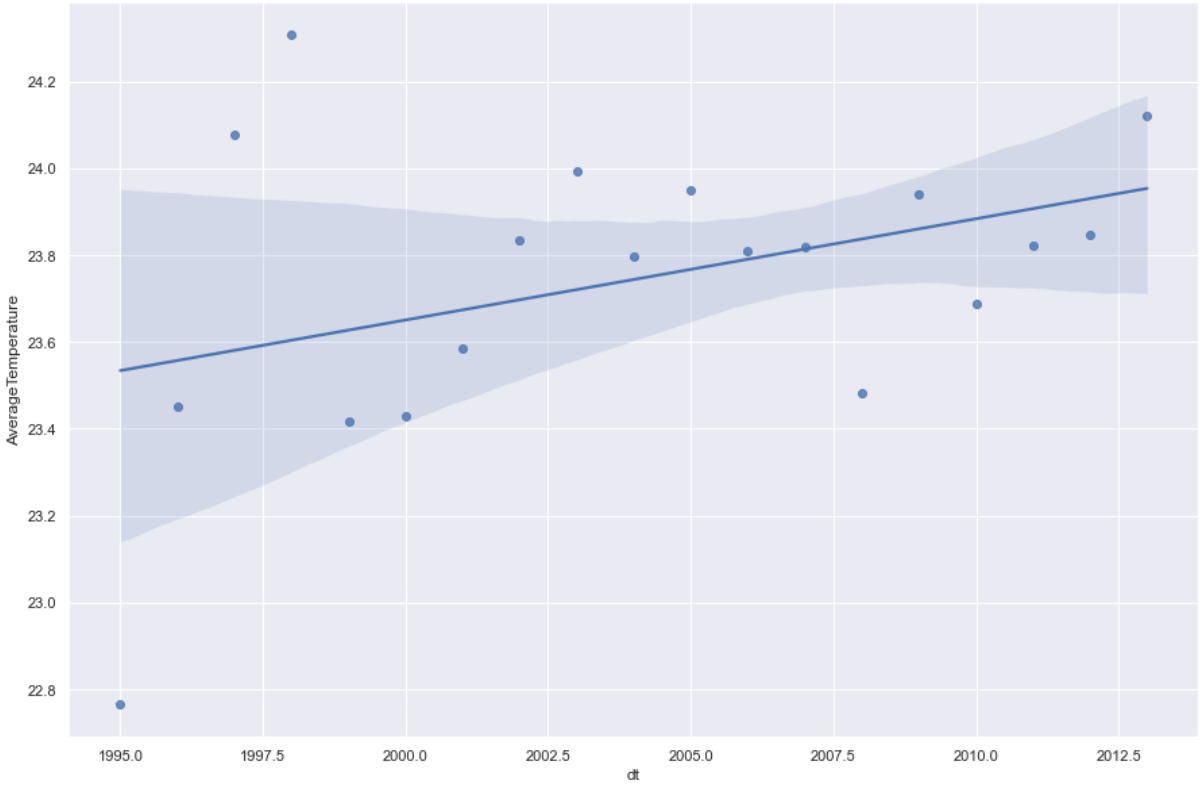
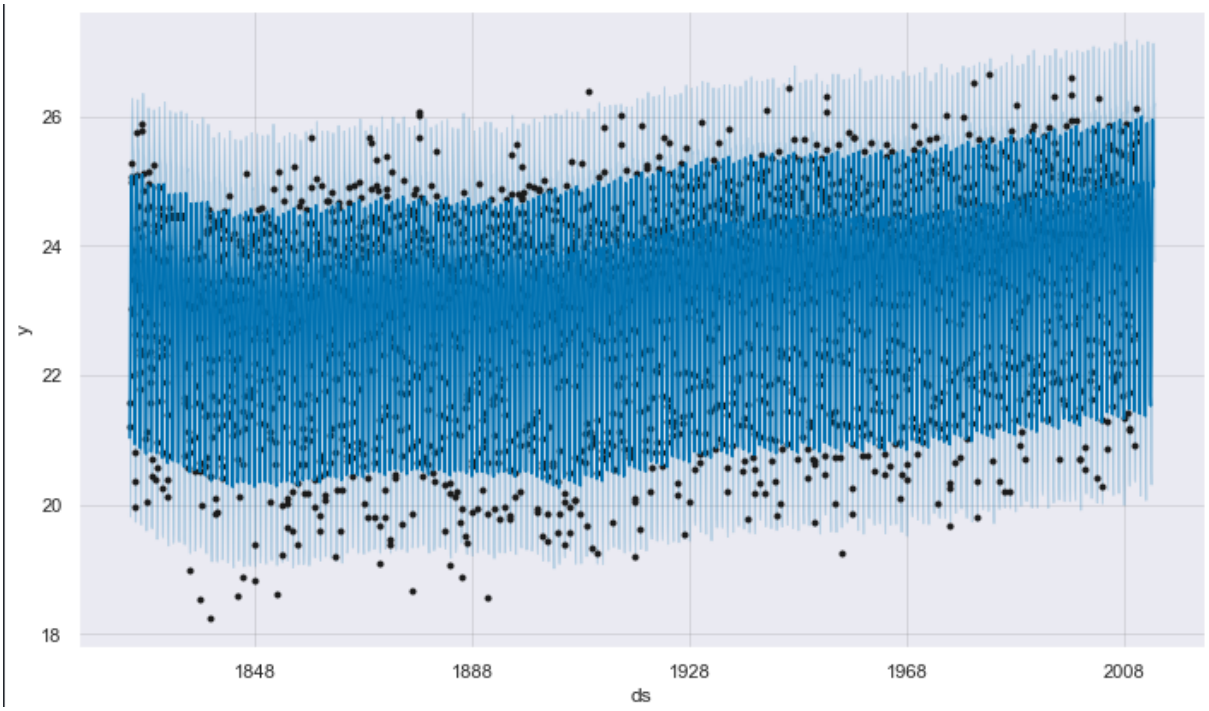
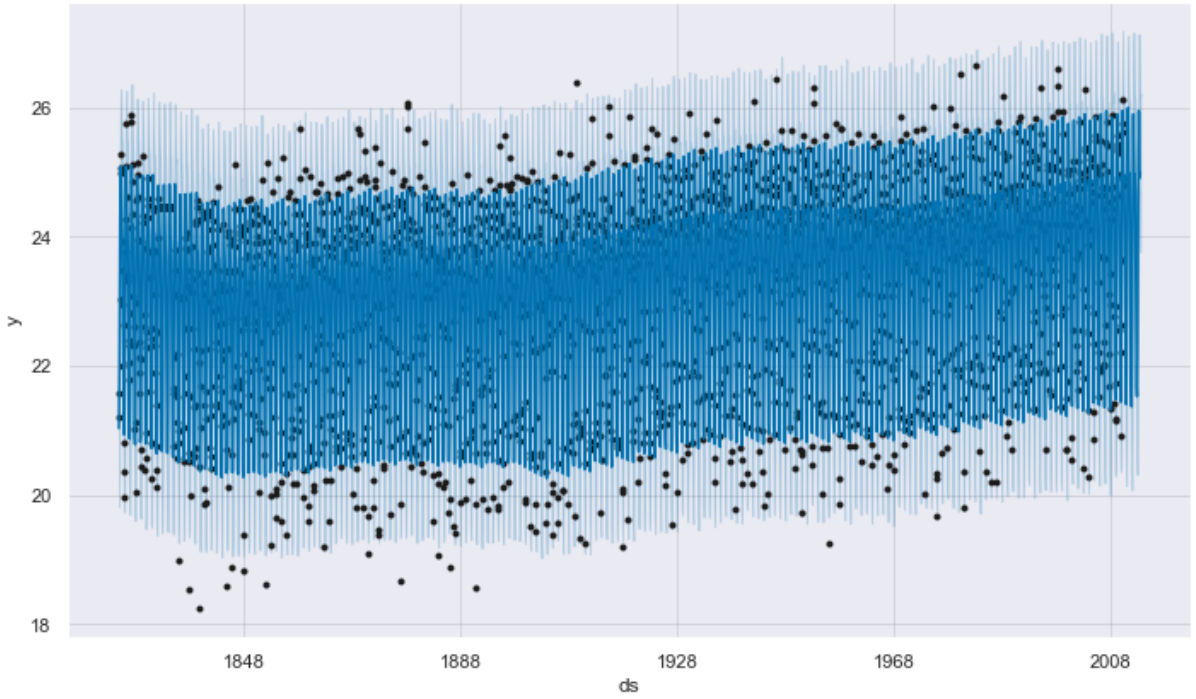


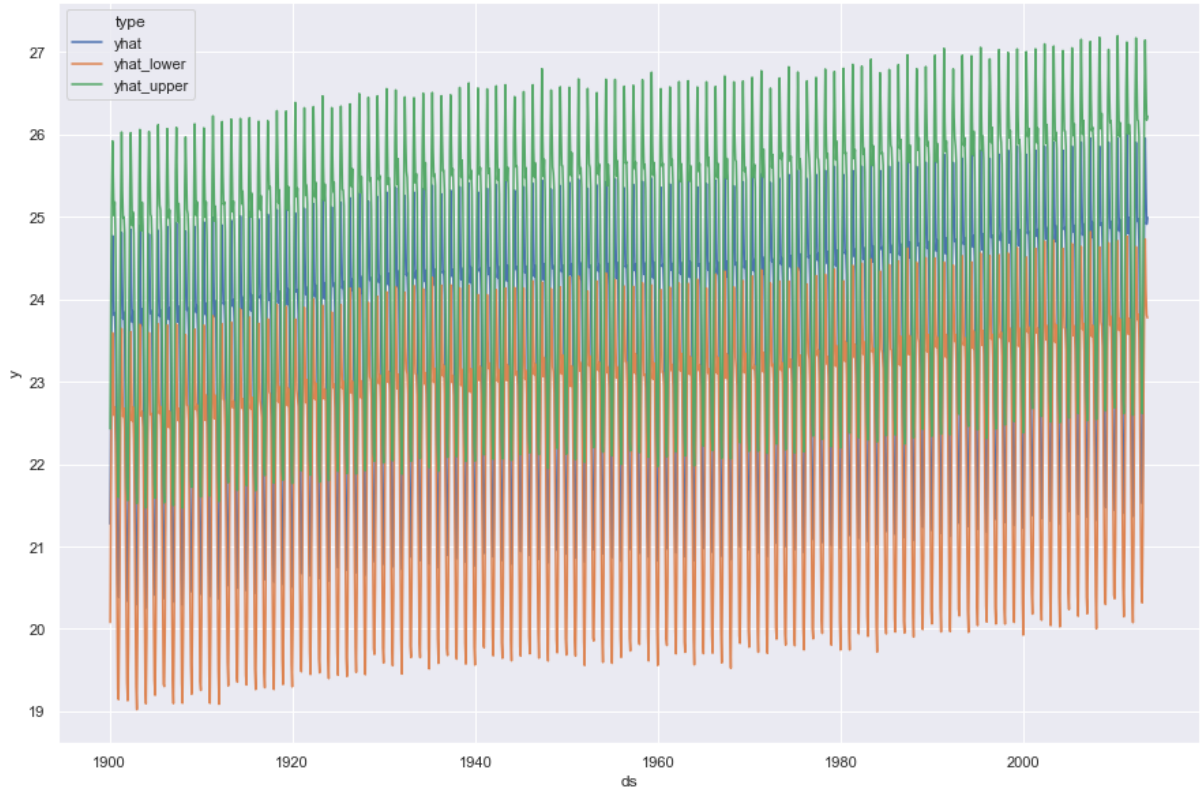
Figura 9: Temperatura promedio en Guatemala en los últimos 20 años (1994 - 2014).

Modelo Prophet para Guatemala

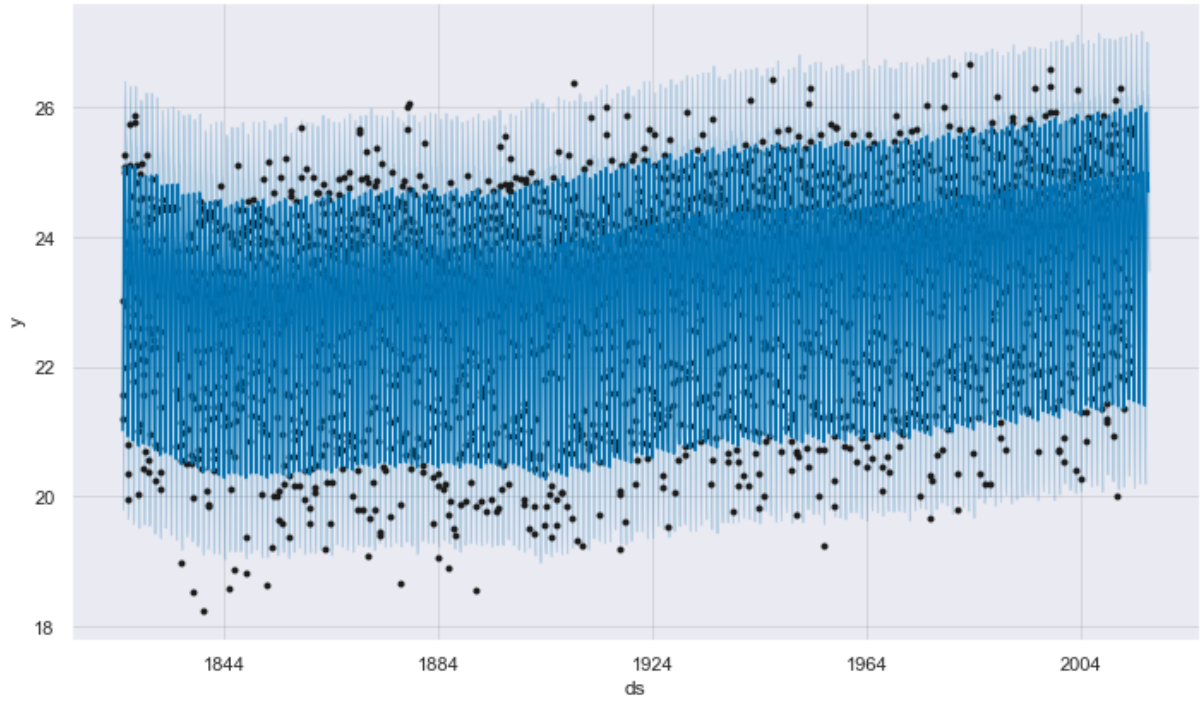


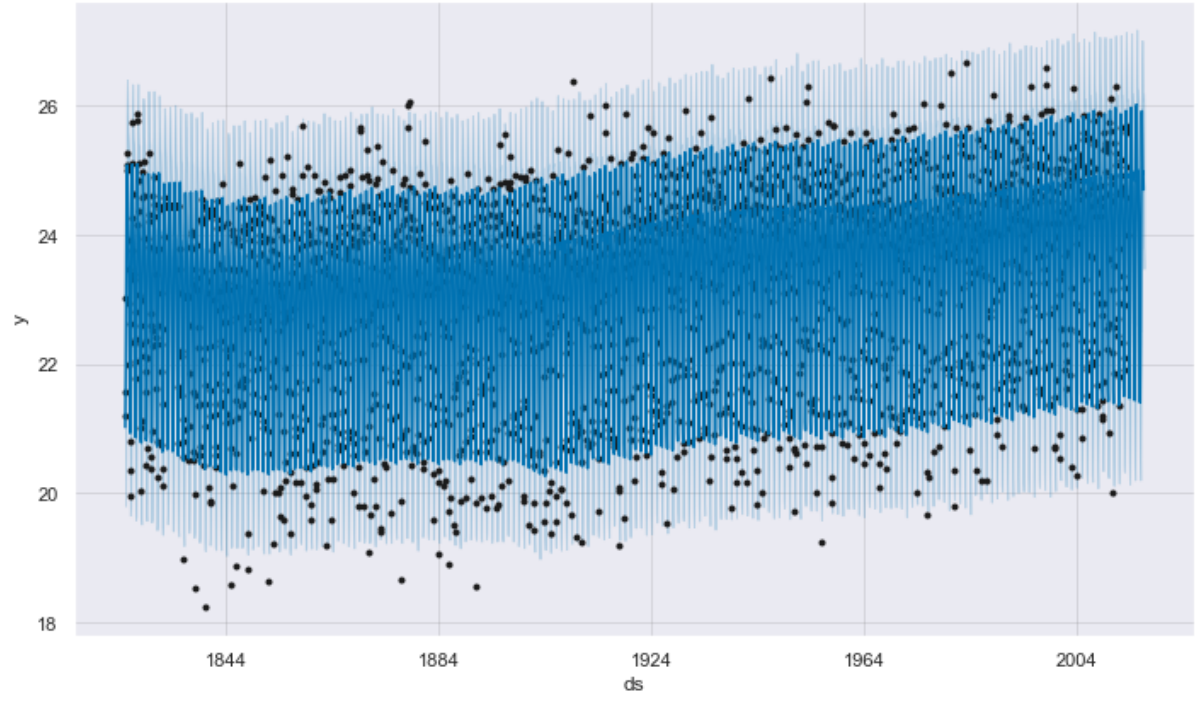


Para Guatemala, se puede observar que las predicciones no son del todo exactas, ya que las gráficas presentan que los valores se traslapan o no se ajustan del todo bien, dando paso a que los datos presenten incertidumbre, sin embargo se obtuvo un valor MAPE de 0.0176 lo que indica un 1.76% de error.

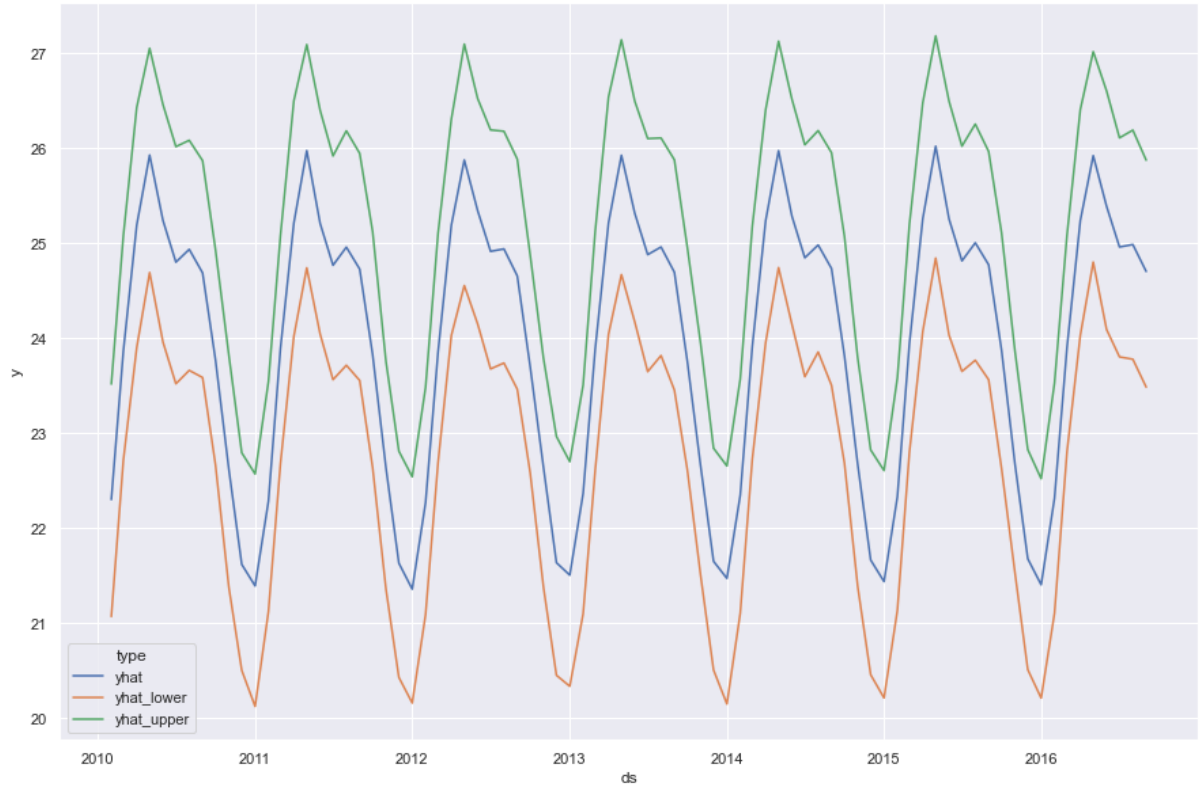


Predicción para Guatemala 3 años a futuro





Para la predicción a 3 años a futuro, podemos observar que los valores de *yhat*, tanto la lower y la upper, se ajustan bastante bien a los datos reales, aunque por la cercanía, también da paso a que ocurran incertidumbres.



El modelo de sarima acertó con la predicción de los datos, sin embargo falló al momento de predecir la tendencia lo cual se hace notorio en las gráficas mostradas, siendo prophet más efectivo por ese lado, pudo predecir de una mejor forma la tendencia de los datos, lo cual lo hizo más efectivo.

Discusión

Cuando hablamos de cambio climático nos referimos al cambio de temperatura y patrones climáticos a largo plazo. Aunque estos cambios pueden ser producidos por efectos naturales existe la consideración de que desde el siglo XIX (años 1801 a 1900) la actividad humana ha aumentado esta variación en la temperatura mundial, dando paso al calentamiento global. Se dice que a partir de la quema de combustibles fósiles como el carbón, petróleo y gas, principalmente con la revolución industrial que se considera del año 1760 al año 1840. Esto se puede evidenciar con la gráfica 11, el componente tendencia de la serie de tiempo que nos muestra un mayor aumento durante el año 1800 y 1840 y luego de esto, sigue aumentando, no tan drásticamente pero sigue aumentando, por lo que consideramos que los datos sí son coherentes con lo indicado por la idea del calentamiento global causado por la actividad humana. (Naciones Unidas, Sin fecha)

Referencias:

NCEI. (2015). Annual 2014 Global Climate Report. Extraído de: https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/201413#:~:text=and%20Climate%20Events-,Global%20Highlights,C%20(0.07%C2%B0F).

NCEI. (2016). Annual 2015 National Climate Report. Extraído de: https://www.ncei.noaa.gov/access/monitoring/monthly-report/national/201513#:~:text=In%202015%2C%20the%20contiguous%20United,temperature%20was%2055.3%C2%B0F.

NCEI. (2017). Annual 2016 Global Climate Report. Extraído de: https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/201613

Naciones Unidas. (Sin fecha). ¿Qué es el cambio climático?. Extraído de: https://www.un.org/es/climatechange/what-is-climate-change